DOCUMENT RESUME

ED 234 629 FL 013 950

AUTHOR Cardoza, Desdemona
TITLE Guidelines for the Evaluation of Bilingual Education Programs.
INSTITUTION National Center for Bilingual Research, Los Alamitos, Calif.
PUB DATE 83
NOTE 33p.
AVAILABLE FROM National Center for Bilingual Research, 4665 Lampson Ave., Los Alamitos, CA 90720.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bilingual Education Programs; Evaluation Methods; Program Effectiveness; *Program Evaluation; Research Methodology; *Research Problems

ABSTRACT

Principles of program evaluation research are outlined so that bilingual education program coordinators can conduct methodologically acceptable program evaluations. The three basic principles of evaluation research are: identification of the program participants, definition of the program intervention, and assessment of program effectiveness. Assessment of program effectiveness requires accurate measurements of how the students performed after program participation and how they would have performed without the program. Among the problems faced in evaluation design is the impossibility of a true experimental design in which limited English speaking students are randomly assigned to bilingual or nonbilingual education programs. In addition, subjects may be from a variety of cultural backgrounds and may enter the program at different ages. Finally, the concept of language proficiency is ambiguous, and a consistent concept and instrument must be used to assess language proficiency among all participants. (RW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED234629

GUIDELINES FOR THE EVALUATION OF BILINGUAL EDUCATION PROGRAMS

Desdemona Cardoza

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

NCBR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

1983

FL013950

national center for bilingual research

SWRL Educational Research and Development
Los Alamitos, California
© 1983

ACKNOWLEDGMENTS

I would like to thank the following individuals for their participation in the workshop: Richard Shavelson, John Nesselroade, Peter Bentler, Gloria Zamora, Norman Gold, Ernest Mazzone, Aurora Quevedo. I would also like to thank Kenyon S. Chan for his invaluable input into these guidelines. The author takes sold responsibility for the contents of this paper.

CONTENTS

Page

Acknowledgments ii

Introduction 1

Three Basic Principles in Evaluation Research 5

Problems in Bilingual Education Evaluation 9

Identification of Program Participations 9

Defining Treatment: Student Participation in Bilingual Education 13

Measuring Program Effectiveness 17

Conclusions 23

Glossary 25

References 27

Appendix A
Workshop Participants 29

Appendix B
Flowchart of the Bilingual Education Evalulation Process 30

INTRODUCTION

This document was prepared for bilingual education program coordinators and all others who are involved with program implementation and evaluation on the local level. Because this group is intimately involved with bilingual education on a daily basis, they will probably all agree that assessing its efficacy is one of the most difficult areas of evaluation research. This is reflected in the lack of methodologically sound evaluations of bilingual education programs (Baker & de Kanter, 1981; Zappert & Cruz, 1977). One of the few attempts at compiling program effectiveness data from official reports submitted by Title VII programs to the funding agency was conducted by Zappert and Cruz (1977). They rejected 97 percent of the reports due to serious methodological shortcomings, including: (a) no control for socioeconomic status, (b) inadequate sample size, improper techniques, or excessive attrition rate, (c) no baseline comparison data, no control group, non-relevant comparison group, (d) no control for initial language dominance, (e) significant differences in teacher qualifications or characteristics, and (f) insufficient statistical information or improper statistical applications.

More recently Baker and de Kanter (1981) examined over 300 documents concerning bilingual education, making it the most comprehensive review to date. The report explores the effectiveness of bilingual education, based on this extensive review of the literature. Of all of the studies initially examined, only 28 were considered adequate methodologically. Baker and de Kanter state that the following characteristics were generally considered the basis for rejection: (a) if students were not randomly assigned to the treatment and comparison groups, and there was no adjustment for possible initial differences between the groups; (b) if the studies did not include appropriate statistical techniques to demonstrate program effectiveness; (c) if the study used the norm-referenced design; (d) if the study lacked a control group and examined gains over the school year; and (e) if the study used grade-equivalent scores.

The Baker and de Kanter piece has received a substantial amount of criticism as well as publicity. The main objection to the report has been that it was somewhat premature. The fact that such a large proportion of the studies reviewed were rejected according to Baker and de Kanter's stringent methodological requirements is indicative of the status of research in the evaluation of bilingual education. This research represents a level of knowledge about the role of language in education that has already been judged as insufficient for policy analysis (Hernandez-Chavez, Llanes, Alvarez & Arvizu, 1982). Bilingual education is in its infancy, and only now are we even approaching the point where we can begin to conduct methodologically sophisticated evaluations. Most practitioners who are involved in the implementation of local bilingual education programs are still learning *how* to do bilingual education. They are aware that there are numerous flaws in their program designs. Their major concern is, therefore, how to improve their programs, or how to design more powerful and effective programs (Piper, 1981). The concern of the federal government is somewhat different. Their major concern is the assessment of program impact. The emphasis on demonstrating program efficacy is, in most cases, premature. Rather than asking the global question of whether a bilingual program works, we should be asking what kinds of programs work best for what kinds of students (Cardenas, 1981).

The "Synthesis of Reported Evaluation and Research Evidence on the Effectiveness of Bilingual Education Basic Projects" (National Center for Bilingual Research, 1982) study is a project being conducted at the NCBR which focuses on assessing the applicability of the meta-analysis procedure in synthesizing the data from bilingual project evaluation reports. Meta-analysis is a methodology for integrating or synthesizing the findings of research. The objectives of this project were to synthesize project evaluations for Title VII Basic Programs and non-Title VII project evaluations as well as other research that focuses on the effectiveness of bilingual education practices with non-English and limited-English speaking students.

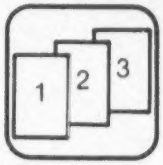
In order for this technique to be successfully employed, it is necessary that certain methodological information be present in the document. For example, outcome data need to be reported and an appropriate comparison group should be used. This technique shows promise in developing and enhancing the knowledge base of bilingual education because it provides a means of reviewing, integrating, and interpreting reports and research studies in the area of evaluation of bilingual education programs. One of the conclusions of the project report was the "The quality of information on project evaluation reports will hopefully be enhanced in the future." Because prior to this time no technique existed for the synthesis and integration of data from Title VII evaluation reports, it is not surprising that the majority of these reports did not meet the criteria for inclusion in the pilot meta-analysis. Because the reports were probably directed at the local school districts, it was not necessary to include some of the information that is critical for data integration at the federal level. The evaluation of bilingual education is an extremely complex area involving numerous legal, ethical, and policy complications. Meta-analysis, along with other strategies, offers a possible method of extending our knowledge and understanding concerning bilingual education practices and their effects.

One of the contributing factors to the problems associated with the research on bilingual education is that, unlike many other federal efforts, this experiment in education was undertaken as an "act of faith" as well as submission to political pressure (Sancho, 1980). There was virtually no existing research upon which to build or refine this educational experiment (Sancho, 1980; Troike, 1978). Over a decade has passed since the passage of the Bilingual Education Act of 1968 (also known as the Title VII amendments to the Elementary and Secondary Education Act of 1965) and, although the quality of research has increased somewhat, it has been surrounded by so much political and academic criticism that results have been declared useless or invalid (Sancho, 1980). Along these lines, Baker and de Kanter's final conclusion is that Title VII bilingual education programs must take steps to improve the quality of their program evaluations.

There are numerous difficulties surrounding the evaluation of bilingual education that research evaluators are probably not aware of, just as there are methodological issues that need to be considered which, in the past, have been ignored by practitioners implementing and evaluating the programs. With this in mind, the NCBR conducted a two-day workshop in an attempt to devise some suggestions for conducting quality evaluations within the constraints of the school and legal systems. The workshop participants included three methodologists with expertise in the areas of longitudinal research, causal modeling, and educational evaluation, and four bilingual education practitioners who were involved with bilingual education program development and implementation at both the state and local level. The purpose of including the four practitioners was to have them identify the problems in the investigation of the impact of bilingual educational services on the educational progress of language minority students. The role of the three research methodologists was to propose potential solutions or alternatives to these problems.

The following discussion is an incorporation of the ideas and recommendations of the workshop participants. This unique combination of individuals represents a long-needed union of knowledge and skills in these two areas. It has produced information which is extremely valuable to researchers in bilingual education evaluation. The NCBR recognizes that the future of bilingual education rests on this issue of properly conducted evaluations. Unless the efficacy of these services can be demonstrated, continuing programs cannot be justified. Because results of past evaluation attempts have been rendered invalid due to inadequate methodology, it is imperative that some type of strategy for examining program effectiveness be developed. This was one of the major goals of the workshop.

THREE BASIC PRINCIPLES IN EVALUATION RESEARCH

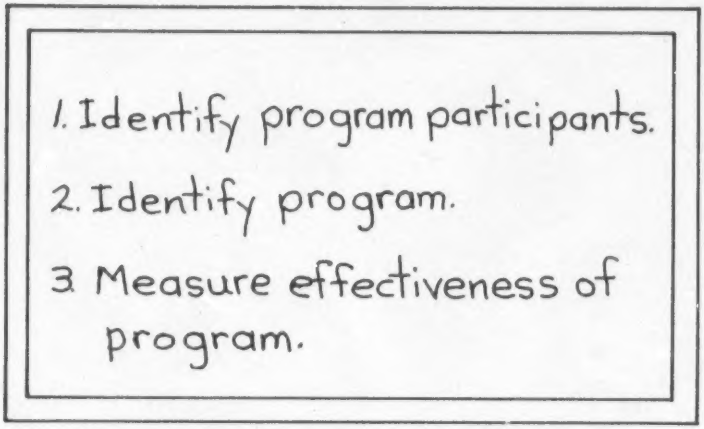


Prior to any program evaluation attempt, the researcher must define who the particular program under consideration is serving. In other words, the program participants must be identified. It is imperative to know who is in the program before any attempt at examining program effectiveness can be made. The second thing the evaluator must know is what *is* the program being examined. How is the treatment defined? It is of relatively little use to demonstrate program effectiveness if one cannot even describe the intervention.

The third principle involves measuring program effectiveness. In order to assess the impact of any program, the evaluator needs two things: (a) A good measure of how the students performed after their program participation; and (b) an accurate estimate of how they would have done without the program. The difference between these measures provides an index by which we gauge project impact. Standardized achievement tests are usually used to assess student performance in the evaluation of bilingual programs. The more difficult problem lies in estimating how language minority students would have done without bilingual education. In other words, the *comparison group* we are really interested in are those students who are limited-English speaking (LES) or non-English speaking (NES) who have remained in the regular classroom where only English is spoken.

The idea would be to conduct a "true" *experiment* using a two-group *experimental design*. An experiment involves comparing the effects of a particular *treatment* with those of a different treatment, or *no treatment*. In a simple, conventional experiment, there is an *experimental group* and a *control group*. The experimental group is exposed to the influence of the factor under consideration; the control group is not. It is then determined what difference appears in the experimental group as contrasted with the control group. Here the assumption is made that the experimental and control groups are equivalent on all factors except the treatment. So in our case, we would assume that the children in the bilingual program differed in no way from those children not in

the program, except for the fact that one group was receiving bilingual services and one was not.



In order to make this assumption, however, subjects would have to be *randomly assigned* to either the experimental or control group. Randomization requires that all individuals in the study be potentially able to participate in either the treatment or no treatment group. In other words, only chance will determine which group an individual is assigned to. This chance assignment assures that there are no *systematic differences* between the groups. It is not enough to present *mean* (i.e., average) differences between the two groups involved and say that this is evidence of program effectiveness. A mere quantitative superiority of the experimental group mean score over the control group mean score is not conclusive proof of its superiority. The means of the two groups are not necessarily going to be identical, even before the treatment is imposed. Any difference that appears at the end of the experimental cycle could possibly be attributed to sampling error. In order to determine whether a difference is *statistically significant,* it is necessary to discredit a sampling error explanation. A commonly used test of significance of the difference between two means is known as a *t-test.* This statistical test involves the computation of the ratio between *experimental variance* (observed difference between two sample means) and *error variance* (the sampling factor).

It is impossible to evaluate a bilingual education program using a true experimental design, however, because students cannot be randomly assigned to one group (receiving bilingual services) or another (not receiving bilingual services). Legally, all NES/LES children must receive *some* treatment, whether it be in the form of a bilingual program or a minimal amount of English tutoring. Therefore, there is no "zero treatment" control group, because all of the children we are interested in are experiencing some type of intervention. It is also not possible to randomly assign students to treatment or control groups. Therefore, in evaluating bilingual education programs, we are limited in design type due to legal restrictions as well as ethical concerns.

One approach to this problem is to use some type of *"quasi-experimental"* design. This refers to experiments that have treatments, *outcome measures,* and experimental units, but do not use random assignment to create the comparisons from which treatment-caused change is inferred (Cook & Campbell, 1979). Instead, *non-equivalent groups* are used for comparison. These groups will probably differ in many ways other than the presence or absence of treatment. What the researchers must do is to try to separate the effects of a treatment from those due to the initial group differences.

The following sections will outline some of these problems which are specific to the evaluation of bilingual education programs, as well as present some of the alternatives and solutions that have been proposed. The key problems as presented by the four bilingual education practitioners are listed in the boxes in the front of each section, followed by a more detailed discussion of each problem as it affects the evaluation process. The alternatives and solutions, as stated by the three research methodologists, appear in the boxes following each section, and are preceded by an explanation specifying how these remedies might be implemented.

PROBLEMS IN BILINGUAL EDUCATION EVALUATION

Identification of Program Participants

PROBLEMS

1. Students are not randomly selected for treatment.
2. Students may be from many different sociolinguistic and cultural backgrounds.
3. Students may enter services at any age.
4. The concept of language proficiency is ambiguous.

The very first step in the evaluation process is to identify who is receiving the services under examination: How was it determined that the students in the program should be there and others should not? The identification of potential candidates provides the first necessary limitation to a full experimental design. Bilingual education services are generally regarded as compensatory services designed to prepare children to adequately benefit from instruction in the regular education programs. Under federal Title VII guidelines, services are designed to provide a transition from the non-English home language to English. With those goals in mind, children are identified as potential candidates for service.

This initial identification procedure is not clearly understood. In some districts all children from language minority backgrounds or from certain ethnic groups are considered potential candidates. These children are all screened at the time of enrollment. Which children are identified for evaluation may vary depending on the ethnic composition of the school, how well the school's personnel adheres to mandated policy, or differences in criteria applied at the state,

district, and school levels. For example, some states, such as Alaska and California, have state-developed screening questionnaires and criteria while other states have no formal process for determining potential candidates for bilingual services. The methodological issue here is that students are not randomly selected from the population for treatment.

Once the pool of candidates is identified, the language proficiency of the child is assessed to ensure that a child is in need of bilingual services. The problem is that there is a general lack of agreement as to what proficiency means. Some districts focus on English proficiency only, without regard to proficiency in the home language. Some regard mastery of basic communicative skills adequate criteria for regular classroom placement, while others require more complex language skills as well. Other problems surface because children may enter programs at any age. Largely due to immigration, students may be identified at any grade level from kindergarten to twelfth grade. Students will also come from many different language backgrounds and cultural groups.

Because students are not randomly selected from the population for treatment (i.e., generally only language minority children are given the services), one must resist the temptation to compare the bilingual education program participants to monolingual English children in regular classrooms. In other words, because we did not randomly select from the general population of students, we must redefine our population to be only NES/LES students, and any comparison group we select must be from this group. The variations in age, sociolinguistic and cultural backgrounds must also be recognized as having potential influence on the outcome measure. These are extraneous variables and must be controlled for. The control of extraneous variables means that the influences of variables irrelevant to the purposes of the study are minimized, nullified, or isolated.

There are a number of ways this can be accomplished. The easiest way, although not always desirable, is to eliminate the variable altogether. In order words, we could use only children from one age group, or children from the same linguistic group, etc. However, by doing this we lose

power of generalization. If we use only one linguistic group, for example, then we can only discuss the program's effectiveness in terms of that particular group.

It may be that we are not interested in any other group. If that is the case, then this is an ideal method of control. If, however, we are interested in more than one linguistic group, then this type of control is probably not the best, and another approach needs to be used. It may be that we want to see if the program has the same effect on several linguistic groups. If that is the case, we can build this variable into the design. This approach can get a bit cumbersome with more than two extraneous variables, and in the case of bilingual education evaluation, there are numerous.

Probably the most effective method is to use a type of statistical control such as analysis of covariance (ANCOVA). This enables the investigator to analyze the difference between groups on the outcome measure, after taking into account initial differences on the extraneous variables. The measure we are controlling for (i.e., age, linguistic group, etc.) is called the covariate. The question we are ultimately interested in is whether the treatment group outperformed the comparison group on the outcome measure by more than should be expected on the basis of initial selection differences. Analysis of covariance provides us with this information.

If a general definition of language proficiency cannot be agreed upon, then at least the concept has to be consistent across all of the students involved in the evaluation. That is, at the very least, the same instrument needs to be used on all study participants to assess language proficiency, and preferably all other testing conditions should be identical (i.e., test time, administrator, etc.)

ALTERNATIVES AND SOLUTIONS

1. Define the population to be all NES/LES students and draw the comparison group from that population.

2. Control extraneous variables by: (a) Eliminating them by looking at only one age group, one linguistic group, etc.; (b) building variables into the experimental design; or (c) using extraneous variables as covariates.

3. Be sure that the definition of language proficiency is consistent across all students in the evaluation.

Defining Treatment: Student Participation in Bilingual Education

PROBLEMS

1. It is not possible to generate discrete program classifications. There is extreme variability in types of programs.
2. Different treatments may occur within a program.
3. There are numerous confounding variables, such as teacher characteristics and parental and community support.

One of the most important questions, but unfortunately also the most difficult to answer, is how to define the treatment. It will probably never be possible to generate discrete program classifications of bilingual education (e.g., maintenance vs. transition). What one individual calls a transitional approach may be very different from what someone else refers to as transitional bilingual education. A more effective procedure would be to try to identify some of the major variables that distinguish different programs. For example, some of the variables that might be important to consider are (a) How do the students learn English? Is it through ESL, concurrent translation, or alternate immersion, etc.? (b) How is the subject matter taught? Is it taught through L_1, or English, or both? If it is taught through both, what is the percentage of time spent using each language? (c) Is there any maintenance of L_1 involved in the program? If so, how is it maintained? (d) Is there a cultural component in the program? If so, how much? (d) What are the reclassification criteria for the program under consideration? (f) What is the length of the services (i.e., one year, three years)?

It is also important to consider within-program variability. Students in a particular program or classroom are typically grouped according to various criteria. It is imperative to identify these criteria. For example, if groups are formed accord-

ing to language proficiency, this is important to know. Groups that are more proficient in English cannot be expected to perform at the same level as those that are less proficient.

In addition to identifying the treatment variables, one must also identify other variables which might be confounding treatment effects. These include teacher variables, parental involvement, administrative and support personnel variables, availability of curricular materials, and non-formal language education.

The difficult task is, of course, measuring these variables. One useful technique would be to use microprogram descriptions. This involves the notion of time-on-task. Time-on-task refers to the time a student spends in a particular content area engaged in learning tasks with a high degree of accuracy. It is based on the premise that in order to examine the impact of classroom practices on student achievement, it is necessary to secure a measure of student learning that can be related directly to classroom phenomena. By using observers in the classroom, information on each student can be collected regarding the amount of time spent on various subjects, in both L_1 and English, how much time the student spends in direct contact with the teacher, what language the teacher is speaking, as well as the language the student is using. By using this technique, an accurate description of the child's classroom experience can be obtained.

When trying to assess teacher characteristics, it would be better to obtain an actual behavioral measure of teacher classroom behavior. The term "behavioral measure" refers to procedures which use systematic observations of teacher classroom behavior to obtain reliable and valid measurements of differences in the typical behaviors which occur in different classrooms, or different situations in the same classroom. A measure is *reliable* to the extent that the average difference between two measurements, independently obtained in the classroom, is smaller than the average difference between two measurements obtained in different classrooms. A measure is *valid* to the extent that it yields scores that reflect actual differences in behavior, and not dif-

ferences in impressions made by different observers. The validity of any behavioral measure depends on three things: (a) A representative sample of behaviors to be measured must be observed; (b) an accurate record of the observed behaviors must be obtained; and (c) the records must be scored so that they accurately reflect differences in behavior.

One behavioral measure that is extremely important to obtain would be a measure of teacher language proficiency in the native language of the students. In addition to using a self-report measure of language proficiency, one could have a native language speaker observe the teacher and then rate his or her proficiency. It would also be desirable to have a behavioral measure of teaching ability which could be obtained in the same manner (i.e., having someone observe the teacher and rate his or her teaching ability).

There are several approaches that could be taken in measuring parent involvement. The frequency of parent committee meetings could be used as an indicator. Another alternative is that the chair of the parent committee meeting could be asked whether or not the parents feel that they have had input into the program. Still another indicator could be whether or not there is a parent training program. Or, finally, one could simply examine the fiscal resources devoted to parent involvement.

Back to the original question of defining treatment: Rather than trying to create separate and distinct categories of bilingual education techniques, a more useful strategy might be to use a cluster analysis. What cluster analysis does is to divide a *set of students* into a small number of relatively homogeneous groups (clusters) on the basis of their similarity over certain variables or attributes (Bailey, 1975). For example, students could be clustered according to their language proficiency and the amount of English instruction they receive. Or they could be grouped according to teacher characteristics and the amount of time they spend learning subject matter in their native language. The purpose here would be to group the students according to their similarity across some of the important variables mentioned above, thereby creating clusters which are more realistic representations of the different types of treatment.

ALTERNATIVES AND SOLUTIONS

1. Use techniques such as time-on-task for program description.
2. Use behavioral measures as much as possible.
3. Use cluster analysis to form different program groups.

PROBLEMS

1. It is difficult to find a comparison group, because no "true" control group can be used.
2. We are often limited in what measures are available for assessing outcome.
3. There are many selection differences between the treatment and comparison group.

Probably the biggest problem which has plagued the investigators in this area is the fact that there can be no "zero treatment" control group. When evaluating the effectiveness of bilingual education, the question we are really interested in is whether the children participating in these services did better than they would have without them. Therefore, the group we are ultimately interested in comparing the treatment group to would be a group of NES/LES children who have received *no* treatment. However, all NES/LES children must receive *some* treatment, to which they are *not* randomly assigned. Therefore, evaluators are restricted to using some type of quasi-experimental design.

Because a true control group cannot be used, it becomes necessary to search out naturally occurring comparison groups. For a variety of reasons there may be children who are eligible for bilingual education services, but who do not receive them. This group of children would be the closest to a control group because they should be similar to the treatment on all important dimensions. This, however, may not be the case. It is important to know why these children were excluded from the services, because this may, in fact, affect the outcome measure for this group. For example, if a district has changed language proficiency measures, some children may have been excluded according to one measure but would have been included by the other. Therefore, according to the second instrument, these children are considered LES, but

are probably more proficient in English than other LES children. This would be extremely important to know when evaluating outcomes for this group.

Another possible comparison group would be those children who are not included in a bilingual program because the district has a policy which requires a minimum number of NES/LES students before a formal program can be instituted. If this is the case, then there may be a small group of children who require bilingual education services but who are, perhaps, only receiving tutoring because there are not enough children to start a formal bilingual program. In this situation, most likely the treatment and comparison groups will come from different schools. When this occurs, then other variables such as socioeconomic status (SES), language background, etc., must be controlled for. Related to this idea is that of using time in treatment per day as a variable. In other words, compare children receiving more treatment per day to those receiving less. But it is important to know why this differential treatment is occurring. If the children are receiving less treatment per day because they are more proficient in English, then this is important to know, and language dominance should be controlled for.

An interesting way to deal with the comparison group problem would be to use a cohort design. The term "cohort" refers to groups of respondents who follow each other through formal or informal institutions (Cook & Campbell, 1979). The use of these cohorts for experimental purposes can be very useful because some cohorts receive a particular treatment, while preceding or following cohorts do not. Oftentimes one can assume that a cohort differs from other cohorts only in minor ways, thereby making an almost ideal comparison group. In the case of bilingual education, the comparison group would be those children who were NES/LES at the time they entered school, but did so prior to the legislation mandating bilingual education. Relevant outcome measures need to be carefully collected on these students who are now presumably in the 8th or 9th grade. This information will then be compared to that collected from the bilingual education participants when they are at the equivalent grade level.

The reason cohort designs are useful for drawing causal inferences is that a "quasi-comparability" can often be assumed between the cohorts that do and do not receive treatment. This, of course, depends on how similar the cohort groups are on the average in terms of certain relevant background characteristics. The assumption is that cohort groups are likely to be more similar to each other than are groups who do not share the same home or school environment. Of course, cohort designs do not automatically rule out selection differences. The design will gain additional strength if it can be demonstrated that cohort groups with and without the treatment do not systematically differ on a third variable that is thought to mediate the treatment effect. For example, when using the cohort design for evaluating bilingual education services, it would be important to establish that the "treatment" cohort and the "no treatment" cohort do not differ significantly on variables such as SES, initial English proficiency, etc.

When selecting outcome measures for the evaluation of bilingual education programs, it would be desirable to use instruments which assess a broad range of outcomes. However, the realistic picture is that we are generally quite limited in what is available. The easiest approach would be to use the district's exit criteria. These might include some measure of language proficiency, a standardized achievement test, parent opinion, and teacher judgment. Quality exit criteria will include all of these indicators; however, a district may use any combination or none at all.

Standardized achievement tests make the most sense as outcome measures because, presumably, one of the main purposes of bilingual education is to give NES/LES children the opportunity to function at the same level as their monolingual English peers in the English-only classroom. It is, however, a good idea to use more than one outcome measure. If we conceptualize a construct of "success," we certainly would like to include academic achievement, but it should also contain other dimensions such as language proficiency in L_2, and also in L_1, if L_1 maintenance is one of the program goals. Teacher evaluation of the student's ability might also prove useful. Measures of self esteem and cultural pride

should also be incorporated if these are goals of the program.

We have now come to the point where it becomes necessary to evaluate differences of the outcome measures across the groups. In other words, was one group more "successful" than the other. Baker and de Kanter (1981) maintain that presenting differences between two groups is not sufficient evidence that the differences did not occur by chance. Appropriate statistical tests must be used to demonstrate program effects.

Because we are using a non-equivalent control group for comparison, some type of adjustment for initial differences between groups needs to be made. A statstical technique that is often used in evaluating program effectiveness is analysis of covariance (ANCOVA). This technique is similar to the analysis of variance (ANOVA) design but also includes one or more covariates in the model to adjust for initial differences between groups. Essentially, what is of interest here is whether the treatment group outperformed the comparison group on the outcome measure(s) by more than would be expected by initial selection differences. So, for example, one might want to use SES and language proficiency as covariates when evaluating the effects of bilingual education, because it is generally thought that these two variables affect academic achievement.

ALTERNATIVES AND SOLUTIONS

1. Search out naturally occurring comparison groups. Use children who (a) should have been in bilingual education, but for some reason were not included; (b) were not included due to a district policy that there be a minimum number of eligible students for program implementation; (c) are receiving a minimum amount of treatment and compare them to those involved in a full bilingual program; (d) should have been eligible for services, but were in school prior to legislation mandating bilingual education. However, initial selection differences must be recognized and statistically controlled for.

2. For outcome measures: (a) Utilize the exit criteria, (b) use multiple indicators whenever possible, and (c) make sure that the outcome criteria match with program content and goals.

3. For initial selection differences, one must (a) identify the variables that the groups differ on that might affect the outcome measure, and (b) use ANCOVA to adjust for these initial differences.

CONCLUSIONS

The purpose of these guidelines is to provide those involved in the evaluation of bilingual education with the necessary basic information needed to conduct technically sound, interpretable evaluation studies. An effort was made to minimize the amount of technical sophistication required in order to benefit from these guidelines. Despite this orientation, program evaluation is an extremely complex area of research and cannot be made simple. Theoretical and computational complexities multiply, particularly when situational constraints force the use of statistical rather than experimental controls for extraneous variables. This is why it may be necessary to seek the consultation of evaluation researchers. These consultants can help with the design of the evaluation, as well as the more complicated aspects such as statistical controls and data analysis. The NCBR recognizes the complexities involved and will assist you in finding the type of help needed. It is important to remember, however, that the responsibility of interpreting the results lies with you as project coordinator and should not be relinquished. Evaluation researchers have training in one area, bilingual education practitioners in another, and both must recognize their limitations. Both must also see that they have significant contributions to make to the area of bilingual education evaluation and that in combination their skills are invaluable.

Finally, it must be recognized that typically very little money in a program's budget is allocated for the evaluation of a bilingual program. This does not mean, however, that a methodologically sound evaluation is impossible. While financial constraints will make the job more difficult, many of the alternatives and solutions offered in these guidelines are relatively inexpensive to execute and, in fact, may prove to be cost efficient. The principles outlined in these guidelines, coupled with creative bilingual and evaluation experts, will improve our understanding of the impact of bilingual programs in the United States.

GLOSSARY

ANCOVA: Analysis of covariance. This technique is used to adjust for initial differences between groups. It extends the ANOVA design by including a covariate in the model in the form of a linear regression.

ANOVA: Analysis of variance. A method for the simultaneous comparison of group means in order to decide if some statistical relation exists between the experimental and dependent variables.

Causal inference: The implication that one variable is causing another.

Cohort design: A quasi-experimental design which utilizes groups which follow each other through formal or informal institutions. One cohort will receive the treatment while preceding or following cohorts will not. Outcome information can be collected from both cohorts and then compared.

Control group: This is the group to which we compare the experimental (treatment) group. In other words, this is the group that receives no treatment.

Covariate: A variable included in the ANCOVA model on which the two groups are thought to be initially different.

Error variance: Differences between groups which result only from the operation of probability or chance.

Experimental group: The group exposed to the influence of the factor under consideration.

Experimental variance: Differences in the outcome measure that may be attributed to the effect of the treatment variable.

Extraneous variables: Those uncontrolled variables (i.e., variables not manipulated by the experimenter) that may have a significant influence upon the dependent variable.

Mean: The mean of a distribution is the arithmetic average.

No treatment group: The same as the control group.

Quasi-experiments: Experiments that have treatments, outcome measures, and experimental units, but do not use random assignment to create the comparisons from which treatment cause change is inferred.

Random assignment: This occurs when an individual has an equal chance of being assigned to either the experimental (treatment) group or the control (comparison) group. What this means is that people with a given idiosyncracy which are assigned to one group will, on the average, be counterbalanced by people with comparable (but not necessarily identical) idiosyncracies in the other group.

Randomization: This involves pure chance selection and assignment of subjects to experimental and control groups from a limited of available subjects.

Selection differences: These differences occur when there is no random assignment of individuals to treatment or control groups. There may be different kinds of people in the two groups.

Time on task: The amount of time that the individual actually spends in overt learning activity.

Treatment group: The same as the experimental group.

Variables: The conditions or characteristics that the experimenter manipulates, controls, or observes.

REFERENCES

Bailey, K. D. Cluster Analysis. In D. R. Heise (Ed.), *Sociological Methodology*. San Francisco: Jossey-Bass, 1975.

Baker, K. A., & de Kanter, A. A. *Effectiveness of bilingual education: A review of the literature.* (Final draft report.) A paper written by the staff of OPBE at the request of the White House Regulatory Analysis and Review Group, September 25, 1981.

Cardenas, J. *Views and voices: An historical perspective.* Institute for Research on Educational Finance and Governance, Vol. 2, No. 4, Autumn 1981. Stanford, CA: Stanford University, 1981.

Cook, T. D., & Campbell, D. T. *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally College Publishing Co., 1979.

Hernandez-Chavez, E., LLanes, J., Alvarez, R., & Arvizu, S. *The federal policy toward language and education: Pendulum or progress?* Cross Cultural Resource Center, Monograph No. 12. Sacramento, CA: California State University, 1982.

National Center for Bilingual Research. *Synthesis of reported evaluation and research evidence on the effectiveness of bilingual education basic projects.* Contract No. 300 81 0439. Los Alamitos, CA: National Center for Bilingual Research, 1982.

Piper, R. *Problems that must be overcome in order to conduct effective evaluations of bilingual education programs.* Draft, NCBR, 1982.

Sancho, A. R. *Bilingual education: A three-year investigation comparing the effects of maintenance and transitional approaches of English language acquisition and aca-*

demic achievement of young bilingual children. Unpublished doctoral dissertation. Claremont, CA: Claremont Graduate School, 1980.

Troike, R. Research evidence for the effectiveness of bilingual education. *NABE Journal,* 1978, *3,* 13-24.

Zappert, L., & Cruz, R. *Bilingual education: An appraisal of empirical research.* Berkeley: BABEL/LAU Center, 1977.

WORKSHOP PARTICIPANTS

Dr. Richard Shavelson
School of Education
University of California
Los Angeles, CA 90024

Dr. John Nesselroade
College of Human Development
Pennsylvania State University
University Park, PA 16802

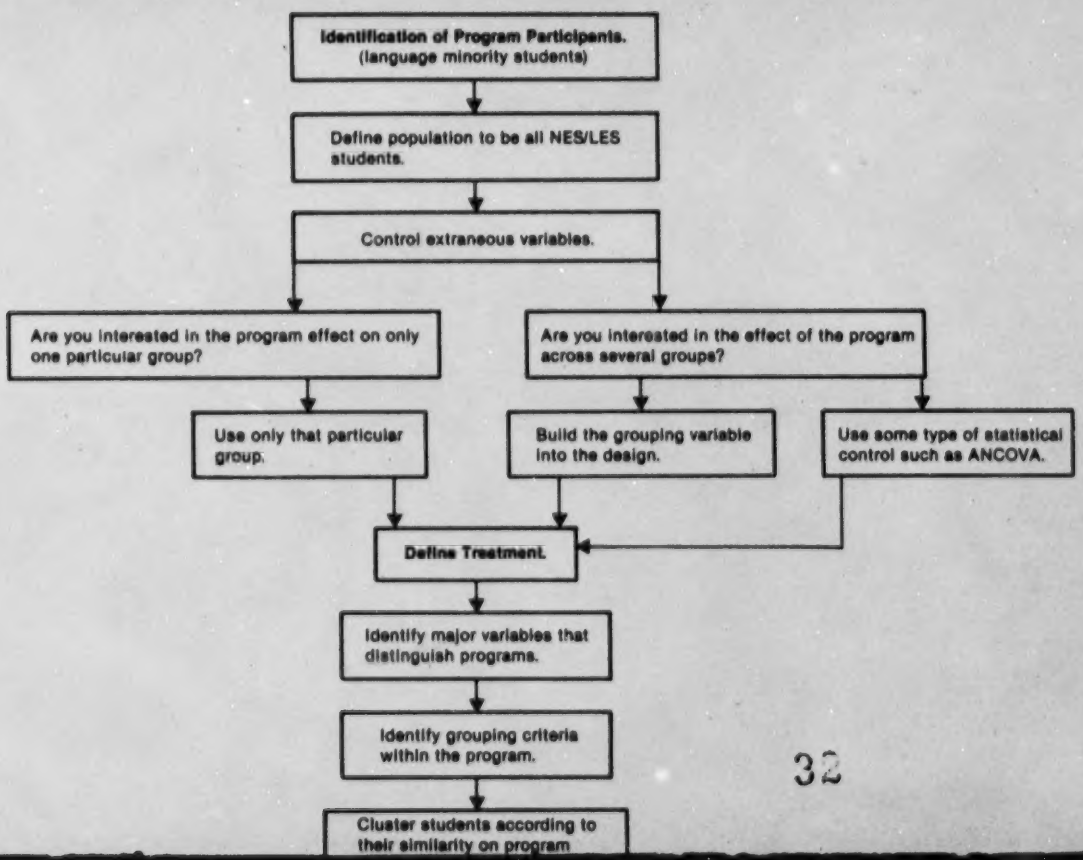
Dr. Peter Bentler
Department of Psychology
University of California
Los Angeles, CA 90024

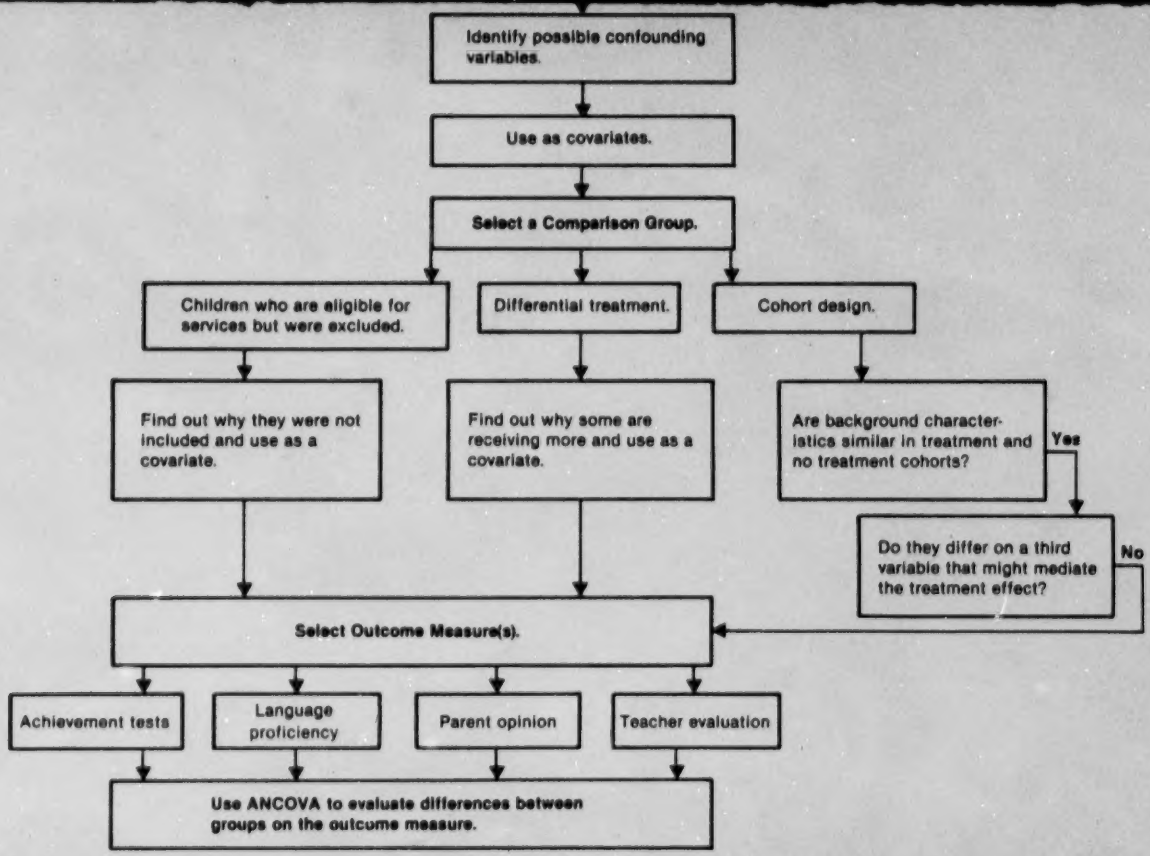
Dr. Glora Zamora
IDRA
5835 Callaghan, Suite 350
San Antonio, TX 78228

Dr. Norman Gold
Office of Bilingual/Bicultural Education
California State Department of Education
721 Capitol Mall
Sacramento, CA 94814

Mr. Ernest Mazzone
Bilingual Education
Massachusetts Department of Education
1386 Hancock Street
Quincy, MA 02169

Ms. Aurora Quevedo
Director of San Jose Area
Bilingual Consortium
1671 Park Avenue
San Jose, CA 95126


Identification of Program Participants. (language minority students)
Define population to be all NES/LES students.
Control extraneous variables.
Are you interested in the program effect on only one particular group?
Are you interested in the effect of the program across several groups?
Use only that particular group.
Build the grouping variable into the design.
Use some type of statistical control such as ANCOVA.
Define Treatment.
Identify major variables that distinguish programs.
Identify grouping criteria within the program.
Cluster students according to their similarity on program


Identify possible confounding variables.
Use as covariates.
Select a Comparison Group.
Children who are eligible for services but were excluded.
Differential treatment.
Cohort design.
Find out why they were not included and use as a covariate.
Find out why some are receiving more and use as a covariate.
Are background characteristics similar in treatment and no treatment cohorts?
Yes
Do they differ on a third variable that might mediate the treatment effect?
No
Select Outcome Measure(s).
Achievement tests
Language proficiency
Parent opinion
Teacher evaluation
Use ANCOVA to evaluate differences between groups on the outcome measure.